Visualização de Dados Estatísticos Representados como Dados Abertos Ligados

Daniele Palazzi[1], Alan Tygel[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.530 – 21941-590 – Rio de Janeiro – RJ - Brasil

danicpalazzi@gmail.com, alantygel@ppgi.ufrj.br

Abstract. *This paper is focused on the analysis of techniques and tools for representation and visualization of statistical data in form of Linked Open Data (LOD). Despite the recent growth in the LOD cloud, the ways of representing statistical data are not yet well founded. The same occurs with the visualization tools, that still can not take profit of the benefits of the representation in triples of the statistical data. In this work, we seek to analyze the particularities of the statistical data, and then review some of the representation forms found in the literature, namely SCOVO, SDMX and Data Cube. We also describe two visualization tools: CubeViz and Visualbox. We conclude that both of them still do not explore all the flexibility and the power of the LOD representation, and give some clues on future work on this area.*

Resumo. *Este trabalho tem o objetivo de analisar técnicas e ferramentas de representação e visualização de dados estatísticos representados na forma de Dados Abertos Ligados (LOD, na sigla em inglês). Apesar do grande crescimento da nuvem de LOD, as formas representar dados estatísticos ainda não estão bem sedimentadas, bem como técnicas de visualização que possam tirar proveito da representação em triplas dos dados estatísticos. Neste trabalho, buscamos analisar as peculiaridades dos dados estatísticos e a partir daí revisar algumas das formas de representação da literatura, como o SCOVO, SMDX e o Data Cube. São descritas ainda duas ferramentas de visualização: Cubeviz e Visualbox. Concluímos que as ferramentas de visualização disponíveis hoje ainda não exploram a flexibilidade e o poder da representação em LOD, e sugerimos alguns caminhos para trabalhos futuros na área.*

1. Motivação

O recente movimento de diversos governos em todo o mundo na direção da criação de plataformas de dados abertos tem impulsionado a pesquisa de formas de representação adequadas para este fim. A expectativa é que estes dados de governos sejam representados em formatos abertos, de fácil acesso à humanos e computadores, e sobretudo, que possa haver integração entre diferentes bases de dados.

Desta forma, podemos dizer que os dados abertos permitem que a democracia se aprofunde no mundo das informações. Cidadãos do mundo inteiro devem ter acesso à informações de governos e empresas, de modo que se possa ter uma ampla transparência e controle social em torno de bens públicos.

Neste sentido, a grande tendência de formato de representação de bases de dados de governo são os dados abertos ligados (LOD, na sigla em inglês: *Linked Open Data*). Através do uso de uso de padrões abertos, suportados pelo W3C, os dados abertos ligados permitem conexões entre bases diferentes, quando uma mesma entidade é representada por um identificador comum em todas a bases que se referirem a essas entidade. Por isso, o formato tem atraído a atenção de órgãos governamentais, dadas as possibilidades de cooperação entre Governo e população através da publicação de dados na Web de uma forma padronizada. Segundo [Berners-Lee, 2006], os princípios de LOD são:

1. Usar URIs como nomes para recursos;

2. Usar URIs HTTP de forma que pessoas possam procurar por estes nomes;

3. Quando alguém procura uma URI, fornecer informação RDF útil;

4. Incluir sentenças RDF que ligam a outras URIs para que possam descobrir outros recursos.

Uma parte significativa dos dados de governo são dados estatísticos. Eles permitem uma leitura geral da sociedade, e a criação de indicadores objetivos com vistas a fazer comparações, traçar metas, ou alertar para algum fato anormal. Apesar do avanço das formas de representação de LOD, a representação dos dados estatísticos ainda não parece muito confortável.

Neste artigo, analisaremos as dificuldades de representar dados estatísticos em forma de LOD e faremos uma revisão dos métodos de representação de dados estatísticos. Analisaremos então as opções de visualização de LOD em geral, e algumas ferramentas de visualização de dados estatístico em LOD, em específico. Em seguida, concluiremos com uma avaliação do estado de desenvolvimento deste campo, e indicaremos algumas possibilidades de avanço.

2. Representação de Dados Estatísticos

Os dados estatísticos podem ser definidos como sequências de observações ao longo de uma ou mais dimensões (tempo, espaço, ou outras). Eles exibem uma peculiaridade em relação, por exemplo, aos dados de uma rede social: são intrinsecamente estruturados. Isso quer dizer que o que faz um dado ser estatístico é que cada observação atende a um conjunto de dimensões e atributos, e é justamente isso que vai nos permitir fazer obervações agregadas e obter dados consolidados. Além disso, os dados estatísticos tendem a existir em grandes quantidades.

Estas características fazem com que a representação de dados estatísticos em triplas não pareça muito adequada. A característica de flexibilidade das triplas não é aproveitada, e acarreta numa enorme redundância de informações, como se pode ver nas Figuras 1 e 2. Não faria sentido, por exemplo, para umas das observações do exemplo

abaixo, haver informação de erro de medida, e para as outras não.

Dia	Temperatura	Cidade
Sábado	37	Rio de Janeiro
Domingo	38	Rio de Janeiro
Segunda	39	Rio de Janeiro
Terça	40	Rio de Janeiro
Quarta	50	Rio de Janeiro

Figura 1: Representação relacional da observação de temperaturas durante cinco dias no município do Rio de Janeiro. Os metadados (dia, temperatura, cidade) são atribuídos de uma vez à toda a coluna, de modo que cada linha já tem seus metadados definidos.

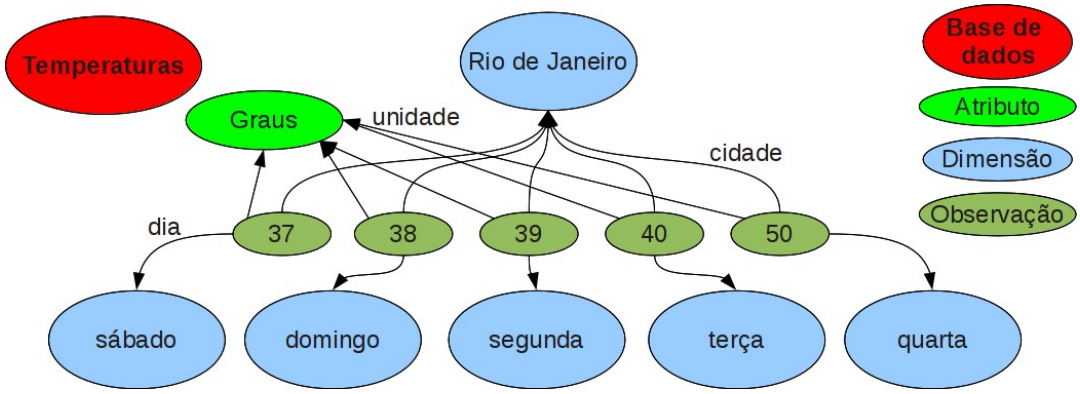

Figura 2: Representação em LOD da medição de temperatura em 5 dias da semana, no município do Rio de Janeiro. Para representar cada observação, são necessárias cerca de 6 triplas, totalizando 30 triplas apenas para as medições.

A seguir, apresentamos algumas das formas de representação dos dados estatísticos e seus meta-dados associados.

2.1. SDMX

Criada em 2001, *Statistical Data and Metadata eXchange* (SDMX[1]), foi motivada pela promoção de padrões para a troca e compartilhamento de informação estatística. Entre as instituições patrocinadoras estão o Banco de Compensações Internacionais (BCI), o Banco Central Europeu (BCE), Eurostat, o Fundo Monetário Internacional (FMI), a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), a Divisão de Estatística das Nações Unidas e do Banco Mundial. Os formatos das mensagens SDMX compreendem duas expressões básicas, SDMX-ML (usando a sintaxe XML) e SDMX-EDI (usando a sintaxe EDIFACT e com base na mensagem estatística GESMES/TS) [Sallas *et. al.* 2012].

1 http://sdmx.org/

2.2. RDF

A fim de permitir uma ampla variedade de aplicações diferentes processar o conteúdo da Web, é importante chegar a acordo sobre formatos de conteúdo padronizados [Heath, T. and Bizer, C. 2011]. Ao publicar dados vinculados na Web, os dados são representados usando o Resource Description Framework (RDF[2]), é uma linguagem de propósito geral para representar a informação na web.

O modelo de dados RDF é projetado para a representação integrada de informações de várias fontes de dados, é heterogeneamente estruturada, e representado usando esquemas diferentes. O modelo é baseado em grafo e possui um alto nível de expressividade.

Os dados são descritos como triplas. As triplas possuem três componentes: sujeito, predicado e objeto. O sujeito é uma URI que identifica o recurso descrito. O objeto pode ser um valor simples ou um literal, ou a URI de um outro recurso relacionado com o sujeito. E o predicado indica que tipo de relação existente entre o sujeito e o objeto, e também é identificado por um URI.

2.3. Statistical Core Vocabulary (SCOVO)

De acordo com [Hausenblas *et. al.* 2009] o Statistical Core Vocabulary (SCOVO) define três conceitos básicos: *dataset*, *data item* e *dimension*.

Um conjunto de dados ou *dataset* representa o contêiner de alguns dos dados, como por exemplo, uma tabela com alguns dados em suas células. Um item de dado ou *data item*, representa um pedaço de dado único, por exemplo, uma célula de uma tabela. A dimensão ou *dimension* representa algum tipo de unidade de um item de dado, por exemplo um período de tempo, localização, etc.

Um conjunto de dados estatísticos em SCOVO é representado por uma classe *Dataset*, que é um conceito SKOS que permite a conexão de um esquema de classificação. Um item de dado estatístico pertence a um conjunto de dados (cf. propriedade inversa *dataset* e *datasetOf*). Um item é subordinado ao conceito de eventos, conforme definido na ontologia de eventos.

A Figura 3 ilustra o vocabulário.

2 http://www.w3.org/TR/rdf-schema/

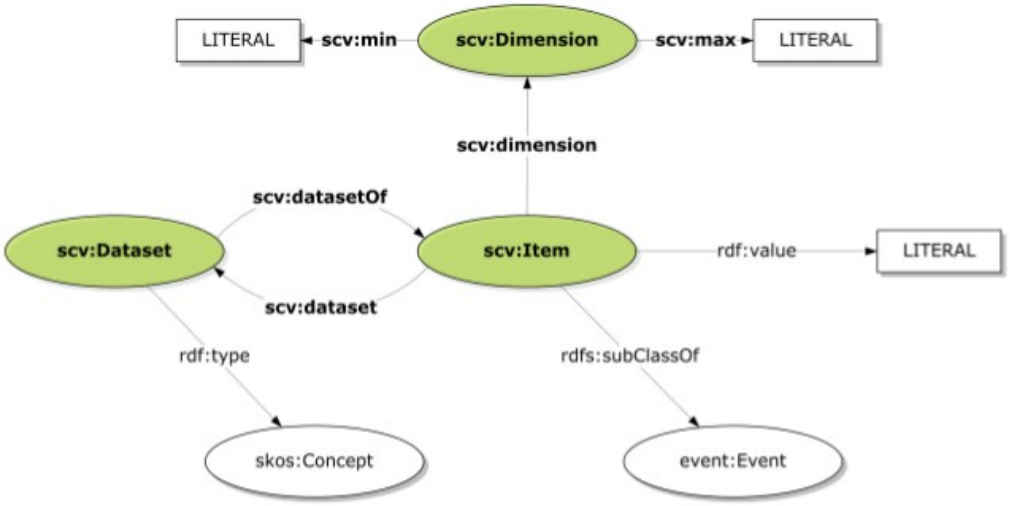

Figura 3: Modelo do SCOVO - Statistical Core Vocabulary.

2.4. Data Cube

O vocabulário Data Cube[3] permite que a informação estastística seja representada usando o padrão RDF e publicada seguindo os princípios de dados ligados. O Data Cube é baseado no padrão SMDX, e é um vocabulário com foco exclusivo na publicação de dados multi-dimensionais na Web.

Segundo [Cyganiak 2010b], a coleção de observações sobre um conjunto de dados estatísticos pode ser caracterizado por um conjunto de dimensões que definem o que a observação aplica-se (por ex., tempo, área, população), juntamente com os metadados descrevendo o que foi medido (por ex., atividade), como foi medido e como as observações são expressas (por ex., unidade de medida).

O modelo de cubo é definido de acordo com um conjunto de dimensões, atributos e medidas. Muito embora utilize-se o termo cubo para representar esses dados, isso não implica que o cubo contem apenas três dimensões, já que o mesmo pode possuir mais e até um número menor de dimensões.

Os componentes de dimensões servem para identificar as observações. Um conjunto de valores para todos os componentes de dimensão é suficiente para identificar uma única observação. Exemplos de dimensões incluem o tempo para o qual a observação se aplica, ou uma região geográfica que abrange a observação.

Os componentes da medida representam o fenômeno a ser observado. E os componentes de atributos permitem qualificar e interpretar o valor observado ou valores observados. Dessa forma, permitem a especificação das unidades de medidas, quaisquer fatores de escala e metadados, como o estado de observação.

Quando se lida com dados estatísticos é usual trabalhar com um subconjunto do grupo de observações de um conjunto de dados. Esse subconjunto recebe o nome de

3 http://www.w3.org/TR/vocab-data-cube/

slice. Eles podem ser um agrupamento útil para que os metadados possam ser ligados, por exemplo, para observar uma alteração no processo de medição que afeta um tempo particular ou região.

3. Visualização de Dados Abertos Ligados

Em [Dadzie, A.-S. & Rowe, M. 2011] são destacados requisitos de alto nível baseados em diretrizes para projetos de informação visual, apresentação, análise e tarefas mais comuns realizadas pelo usuário, para o consumo de dados ligados. De acordo com os autores, as ferramentas para visualização devem: apresentar uma navegação intuitiva; explorar dados de modo a obter uma compreensão de sua estrutura e conteúdo; identificar as ligações dos conjuntos de dados; identificar os erros e outras anomalias no conteúdo e na sintaxe; permitir consultas avançadas; publicação e distribuição; e permitir a extração de dados para reutilização em outras aplicações.

Tradicionalmente, dados abertos ligados são visualizados sob a forma de um grafo. Existem hoje inúmeros navegadores que permitem a navegação entre fontes de dados expressas em triplas RDF. Alguns deles são baseados apenas em textos, e outros permitem uma melhor análise visual com várias formas de apresentação. Dentre as diversas ferramentas disponíveis, foram selecionadas duas para uma análise mais detalhada.

O *Disco Hyperdata Browser*[4] é um navegador simples que processa todas as informações que podem ser encontradas sobre um recurso específico, como por exemplo, uma página HTML. Estes recursos, por sua vez, contêm *links* que possibilitam a navegação entre os recursos. Para começar a navegar, o usuário insere uma URI na caixa de navegação. Ao pressionar o botão "Go!", o navegador obtém informações sobre o recurso desejado e a informação é recuperada, exibindo assim uma tabela contendo propriedades, valores e fontes. As fontes que contêm as informações específicas são referenciadas pelas siglas G_1,…, G_n. Enquanto estamos navegando de um recurso para o outro, o navegador armazena todos os gráficos RDF recuperados em um cache de sessão. Ao clicar sobre "Display all RDF graphs" abre-se uma nova janela com a lista de todos os gráficos RDF recuperados e todas as URIs que não foram dereferenciadas com sucesso. A Figura 4 mostra a interface do navegador.

4 http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/disco/

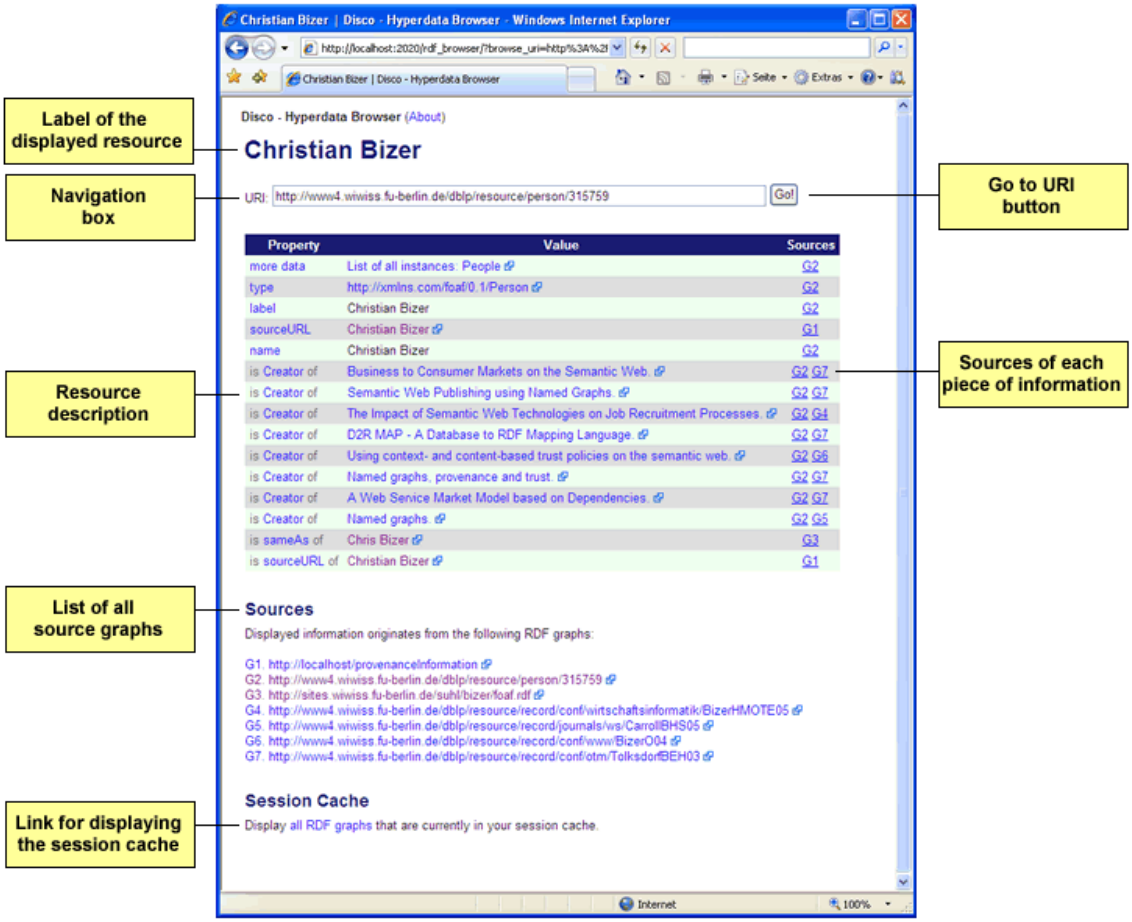

Figura 4: Interface *Disco Hyperdata Browser.*

A *RDF GRAph Visualization Tool* (*RDF Gravity*)[5] é uma ferramenta para visualização de grafos construídos em RDF ou OWL, que fornece uma visualização simples, mas poderosa de gráficos RDF, e a capacidade de filtrar e visualizar partes específicas de um gráfico. As características principais são: visualização gráfica (renderização, zoom, seleção); filtros (global, local, *namespace* em nível de instância); busca de texto e consultas RDQL; e visualização de múltiplos arquivos RDF. A Figura 5 exibe a ferramenta RDF Gravity.

5 http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html

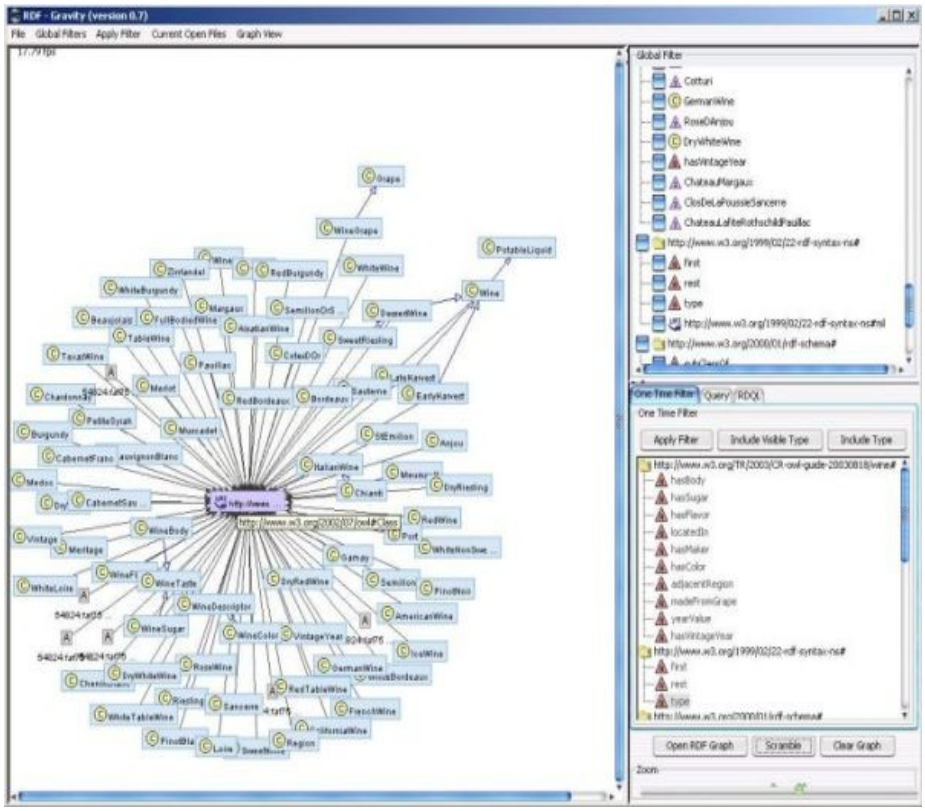

Figura 5: Interface RDF-Gravity.

4. Visualização de dados estatísticos representados como LOD

A princípio, a forma de visualização de um determinado conjunto de dados é independente da sua forma de representação. Afinal, se temos a informação, seja qual for a forma de representação, podemos mostrar esta informação da maneira mais conveniente.

Entretanto, na prática, a forma de representação pode ter uma grande influência nas possibilidades de uso dos dados, inclusive na visualização. As ferramentas de visualização são construídas a partir das representações, e exploram as facilidades que elas oferecem. Também ficam sujeitas às limitações de cada forma de representação.

A visualização de dados estatísticos é uma área com vários anos de estudo, e sua revisão está fora do escopo deste trabalho. Citamos apenas algumas ferramentas hoje disponíveis de forma *online*, onde se pode explorar a riqueza de várias formas de visualização aliadas à boas e intuitivas opções de configurações. O portal Google Public Data[6] disponibiliza diversas bases de dados oficiais, e ainda permite que dados dos usuários sejam publicados. O formato utilizado é o DSPL, através do qual se definem as observações, dimensões, atributos e cortes (*slices*) que serão aplicados aos arquivos CSV contendo os dados. Outro portal que oferece serviço semelhante é o Manyeyes[7].

Para visualizar dados estatísticos representados em LOD, o campo de opções de reflexões teóricas e ferramentas ainda é bastante estreito. Em [Fernéandez *et. al.* 2012]

6 http://www.google.com/publicdata
7 http://www-958.ibm.com/software/analytics/manyeyes/

temos a proposta de um modelo genérico de visualização de dados em LOD. Em [Salas *et. al.* 2011] temos a descrição de um *workflow* para publicação de dados estatísticos em LOD, que no entanto passa de forma rápida pela visualização, através de uma das ferramentas (CubeViz) que será descrita a seguir. Em [Sthur *et. al.* 2011] é apresentada a ferramenta LODWheel, que parece ter caraterísticas interessantes relacionadas aos dados estatísticos. Entretanto, o caminho apontado para a ferramenta não está funcional. [Zapilko *et. al.* 2011] propõe um modelo de trabalho com dados estatísticos em LOD a partir de diferentes bases de dados e partir de combinações entre dados estruturados e triplas, procurando explorar o melhor de cada representação. Em outro artigo [Zapilko e Mathiak 2011], os autores propõem um modelo de testes de qualidade com vistas à integração entre bases de dados estatísticos em LOD.

Das ferramentas encontradas, consideramos que duas delas têm potencial para maior desenvolvimento e utilização. A Tabela 1 mostra a comparação entre as duas ferramentas, e as subseções a seguir descrevem cada uma.

Tabela 1: Comparação entre as ferramentas

	CubeViz	**Visualbox**
Linguagem de programação	PHP	PHP
Ambiente	Plugin para Ontowiki	Standalone
Quem desenvolve	AKSW (Alemanha)	Álvaro Graves (Chile)
Representação dos dados	Triplas RDF / Datacube	Triplas RDF
Desenvolvimento	http://aksw.org/Projects/CubeViz.html	http://alangrafu.github.com/visualbox/
Exemplos	http://cubeviz.aksw.org/	http://visualbox.org/demo/index.html
Opções de visualização	Tabela, Pizza, Barras (Horizontal e Vertical), Radial, Linhas, Montanha	Tabela, Pizza, Barras (Horizontal e Vertical), Radial, Linhas, Nós Temporais, Mapas, Nuvem de Palavras, Coordenadas paralelas, Grafos, Dendograma, Circulos hierárquicos

Licença	GPL, com componentes Apache License e Creative Commons Attribution-NonCommercial 3.0	Apache License

4.1. CubeViz

O CubeViz foi desenvolvido pelo *Agile Knowledge Engineering and Semantic Web* (AKSW), da Alemanha, no âmbito do projeto LOD2. Funciona como um *plugin* do Ontowiki, ferramenta de navegação por LOD desenvolvida pelo mesmo grupo.

O CubeViz funciona em conjunto com o Stats2RDF, outro *plugin* que converte arquivos CSV para triplas representadas usando o Data Cube, por um processo semiautomático.

Após a importação dos dados, a ferramenta permite a visualização facetada em gráficos e tabelas predefinidos. A partir das dimensões detectadas, é possível fixar algumas e escolher duas ou mais instâncias das outras dimensões para comparação.

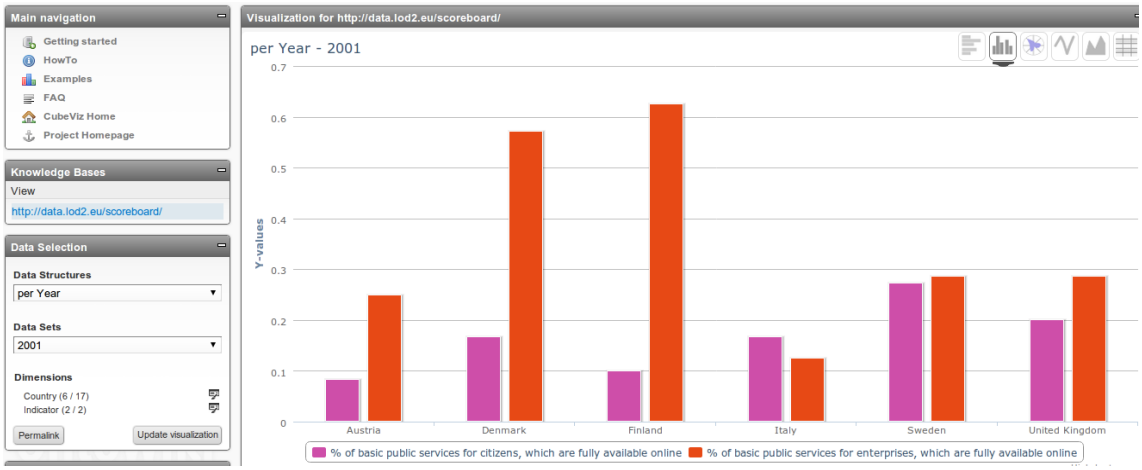

Figura 6: Tela de Visualização do CubeViz. A esquerda abaixo, podemos ver a seleção das dimensões. O gráfico mostra a comparação de dois indicadores (rosa e vermelho) entre 6 países, para o ano de 2001.

4.2. Visualbox

O Visualbox é uma versão simplificada do LODSPeaKr, feita por um desenvolvedor chileno (Alvaro Graves). É uma ferramenta genérica de visualização de dados em LOD a partir de um endpoint qualquer

Seu workflow simples permite utilizar uma base qualquer e gerar a visualização desejada. A partir de 3 ações é possível gerar uma visualização:

1 – É necessário escolher um e*ndpoint*, ou seja, uma base de dados em RDF acessível. O *endpoint* pode ser local ou remoto.

2 – É necessário construir uma consulta SPARQL para recuperar os dados que

serão mostrados na visualização.

3 – A partir de opções de visualização dos dados – grafo, gráficos, tabelas – cria-se um *template* HTML que pode ser usado sozinho ou embutido em alguma aplicação.

Figura 7: Tela de Configuração da Visualização do Visualbox

4.3. Dificuldades de Instalação

Apesar das perspectivas interessantes, não foi possível instalar e utilizar completamente nenhum dos dois softwares. A utilização e comparação foi feita utilizando as versões parcialmente instaladas localmente, e as versões de demonstração fornecidas pelos desenvolvedores.

No caso do Visualbox, a instalação foi feita normalmente utilizando a plataforma de controle de versão GIT. É possível acessar a página e visualizar a tela mostrada na Figura 7. Ocorrem então dois problemas: (1) O acesso ao *endpoint* local (Sesame) é negado. Assim, o teste de consulta SPARQL retorna um erro de permissão de acesso, e a visualização do componente falha por não ter dados para mostrar. (2) O acesso a *endpoints* externo funciona, e as consultas de teste para bases como a DBPedia retornam resultados normalmente. Entretanto, quando se tenta visualizar o componente, os dados não são enviados, e não é possível ter o resultado esperado.

Em relação ao Cubeviz, o maior problema foi a instalação de um componente de abstração de banco de dados chamado ODBC. Os manuais de instalação não mencionam este componente, entretanto ele é necessário para que o Virtuoso se conecte com o servidor PHP.

5. Conclusões

É importante ressaltar que o campo retratado neste trabalho – visualização de dados estatísticos em forma de dados aberto ligados – é um campo recente com pouco trabalho ainda desenvolvido. Apesar dos grandes avanços vistos recentes na visualização de dados abertos ligados em geral, ainda não foi dada a devida atenção aos dados estatísticos.

Como dito anteriormente, a forma de visualização de um dado não é necessariamente dependente da sua forma de representação. Por isso, esperamos que, no mínimo, seja possível ter os mesmo tipos de visualização a que estamos acostumados para dados abertos ligados, como todos os tipos gráficos, tabelas, cartogramas etc.

No entanto, a representação em dados abertos ligados oferece uma gama de outras possibilidades a partir de três características: flexibilidade de representação, ligações e semântica. Estas características devem ser exploradas para que possamos ter formas de visualização que possam receber dados de várias fontes, sejam mais interativas através das possibilidades de ligações, e que possam se adaptar a um determinado contexto através do sentido semântico que os dados podem adquirir. Isto, é claro, sem abrir mão da eficiência que temos no modelo relacional para tratamento de grandes quantidades de dados.

Neste trabalho, identificamos diversos avanços na área. O primeiro deles é o formato de representação Data Cube, que adapta o estado da arte em metadados estatísticos (SDMX) ao mundo LOD.

Além disso, temos hoje disponíveis uma ferramenta altamente genérica (Visualbox) e outra integrada a um potente *framework* de navegação em triplas (Cubeviz/Ontowiki), que apontam os caminhos a serem seguidos: ferramentas que exploram a flexibilidade, e ferramentas que tirem proveito das ligações e da semântica.

À medida em que as ferramentas forem se tornando mais utilizadas, teremos certamente avanços nas facilidades de instalação e utilização. Usá-las hoje ainda é um desafio de habilidades computacionais, devido à pequena comunidade e pouca documentação.

Mesmo assim, acreditamos que em breve seja possível utilizar todo o potencial dos dados abertos ligados para a construção de visualizações criativas que possam de fato atender o objetivo final: transformar grandes quantidades de dados em informação útil, e transformar essa informação em conhecimento.

Bibliografia

Berners-Lee, T. (2006). Linked Data - Design Issues. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html

Cyganiak R, Field S, Gregory A, Halb W and Tennison J (2010), "Semantic Statistics: Bringing Together SDMX and SCOVO", In LDOW. Vol. 628 CEUR-WS.org.

Cyganiak R, Reynolds D and Tennison J (2010), "The RDF Data Cube vocabulary"

Dadzie A-S and Rowe M (2011), "Approaches to Visualising Linked Data: A Survey",

Semantic Web 1.

Fernéandez JMB, Auer S and Garcia R (2012), "The Linked Data Visualization Model.", In International Semantic Web Conference (Posters & Demos). Vol. 914 CEUR-WS.org.

Hausenblas M, Halb W, Raimond Y, Feigenbaum L and Ayers D (2009), "SCOVO: Using Statistics on the Web of Data", In Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications. Berlin, Heidelberg , pp. 708–722. Springer-Verlag.

Heath T and Bizer C (2011), "Linked Data: Evolving the Web into a Global Data Space" Morgan & Claypool.

Salas PE, Martin M, Mota FMD, Breitman K, Auer S and Casanova MA (2012), "Publishing Statistical Data on the Web", In Proceedings of 6th International IEEE Conference on Semantic Computing. Palermo, Italy IEEE.

Stuhr M, Roman D and Norheim D (2011), "LODWheel – JavaScript-based Visualization of RDF Data", In Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011),., October, 2011.

Zapilko B, Harth A and Mathiak B (2011), "Enriching and Analysing Statistics with Linked Open Data", Eurostat Conf. on New Techniques and Technologies for Statistics (NTTS).

Zapilko B and Mathiak B (2011), "Defining and Executing Assessment Tests on Linked Data for Statistical Analysis.", In COLD. Vol. 782 CEUR-WS.org.